BS6200 Project Report

M. Sc. in Biomedical Data Science

Feb 2020

Han Runbing

Table of Contents

1. Introduction

1.1 Project Background

Dengue fever is a mosquito-borne tropical disease caused by the dengue virus. Symptoms typically begin three to fourteen days after infection [1]. These may include a high fever, headache, vomiting, muscle and joint pains, and a characteristic skin rash.

In recent years dengue fever has been spreading. Since it is carried by mosquitoes, the transmission dynamics of dengue are related to climate variables such as temperature and precipitation. An understanding of the relationship between climate and dengue dynamics can improve research initiatives and resource allocation to help fight life-threatening pandemics.

Historically, the disease has been most prevalent in Southeast Asia and the Pacific islands. For this project, there are two cities, San Juan and Iquitos, with test data for each city spanning 5 and 3 years respectively. Given the data of dengue cases and the climate of those two cities, the goal of this project is to predict the number of dengue cases each week (in each location) based on environmental variables describing changes in temperature, precipitation, vegetation, and more.

An understanding of the relationship between climate and dengue dynamics can improve research initiatives and resource allocation to help fight life-threatening pandemics.

1.2 Introduction of Dataset

1.2.1 Data Sources

The data of this project comes from multiple sources. Dengue surveillance data is provided by the U.S. Centers for Disease Control and prevention, as well as the Department of Defense's Naval Medical Research Unit 6 and the Armed Forces Health Surveillance Center, in collaboration with the Peruvian government and U.S. universities. Environmental and climate data is provided by the National Oceanic and Atmospheric Administration (NOAA), an agency of the U.S. Department of Commerce.

1.2.2 Features

- City and Date Indicators

city – City abbreviations: *sj* for San Juan and *iq* for Iquitos

year – Year of each record

weekofyear – The serial number of weeks in each recording year

week_start_date – Date given in yyyy-mm-dd format

- Temperature Related Features

station_max_temp_c – Maximum temperature

station_min_temp_c – Minimum temperature

station_avg_temp_c – Average temperature

station_diur_temp_rng_c – Diurnal temperature range

reanalysis_dew_point_temp_k – Mean dew point temperature

reanalysis_air_temp_k – Mean air temperature

reanalysis_max_air_temp_k – Maximum air temperature

reanalysis_min_air_temp_k – Minimum air temperature

reanalysis_avg_temp_k – Average air temperature

reanalysis_tdtr_k – Diurnal temperature range

- Precipitation Related Features

station_precip_mm – Total precipitation

precipitation_amt_mm – Total precipitation

reanalysis_sat_precip_amt_mm – Total precipitation

reanalysis_precip_amt_kg_per_m2 – Total precipitation

- Humidity Related Features

reanalysis_relative_humidity_percent – Mean relative humidity

reanalysis_specific_humidity_g_per_kg – Mean specific humidity

- Normalized difference vegetation index (NDVI)

ndvi_se – Pixel southeast of city centroid

ndvi_sw – Pixel southwest of city centroid

ndvi_ne – Pixel northeast of city centroid

ndvi_nw – Pixel northwest of city centroid

1.3 Introduction of Performance metric

- Mean Absolute Error (MAE)

MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It is the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight[2].

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right|$$

- Mean Squared Error (MSE)

MSE is a quadratic scoring rule that also measures the average magnitude of the error. It's the square of the average of squared differences between prediction and actual observation.

$$MSE = \frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2$$

2.Methodology

2.1 Exploratory Data Analysis

To have a deep insight into the dataset, the first analysis is to explore and visualize the patterns of class label and features. After analysis, the significant difference between two cities is discovered. That is why the further process should be applied to two cities respectively. Besides, the relatively correlation between cases time as well as the uncorrelation between cases and other features are also indicated from the analysis.

2.1.1 Time Series Analysis

From the whole pattern cases as time series value, there are several dominate peaks in two cities, which indicates the break of dengue may correlate with time **Fig 1a**. From the scatter of week of year and total cases, the seasonal change can be discovered in both cities **Fig 1b**. To be more specific, the impact of seasonal change is larger in San Juan. **Fig 1c**.

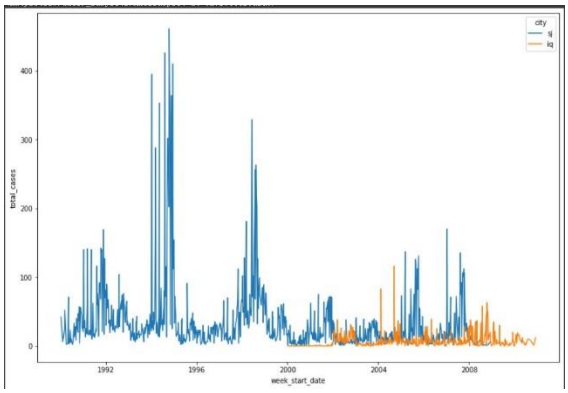
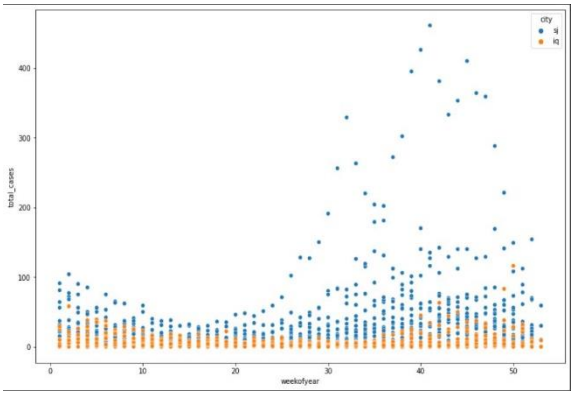


Figure 1a. The Whole Trends of Cases



1b. The scatter of cases and weeks

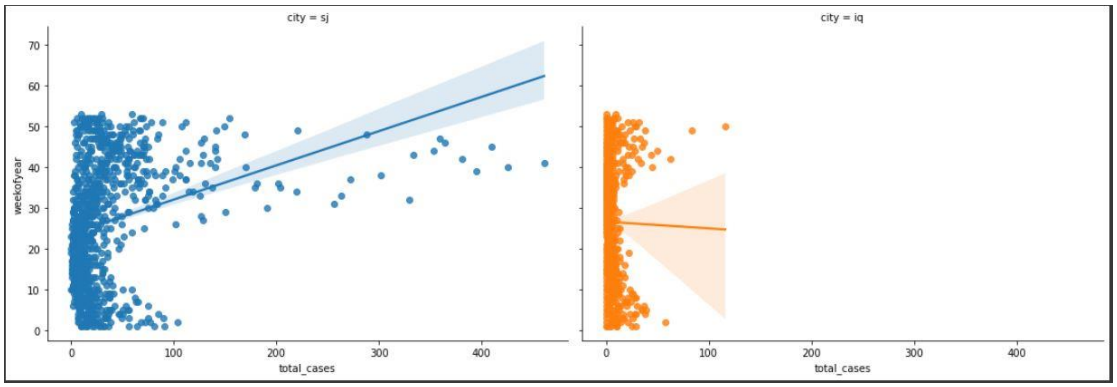


Figure 1c. The correlation of total cases and weeks

2.1.2 Difference Between Cities

- Sample number

The first different between cities is the sample size **Fig 2**. There are two cities, San Juan and Iquitos, with test data for each city spanning 5 and 3 years respectively.

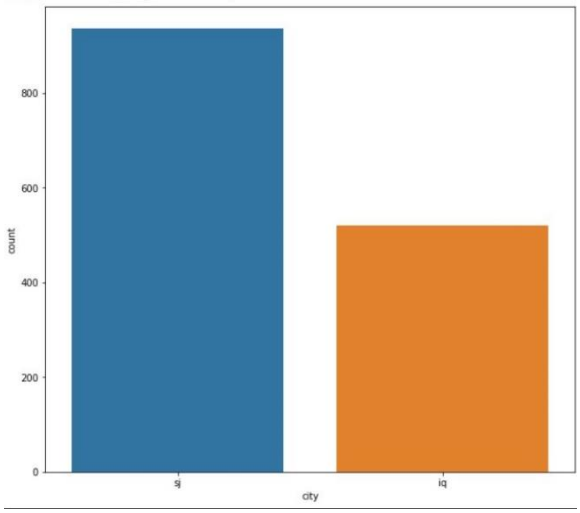
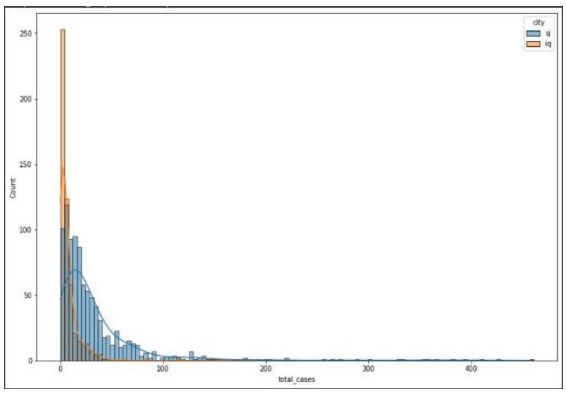


Figure 2. Difference of Data Size Between two cities

- Difference of Distribution

From histogram of cases in two cities, the number of total cases in San Juan trends to be higher the that in Iquitos **Fig 3a**. Furthermore, from the violin distribution, there are some extremally high cases shown in San Juan **Fig 3b**, which contributes to the large difference between two cities.



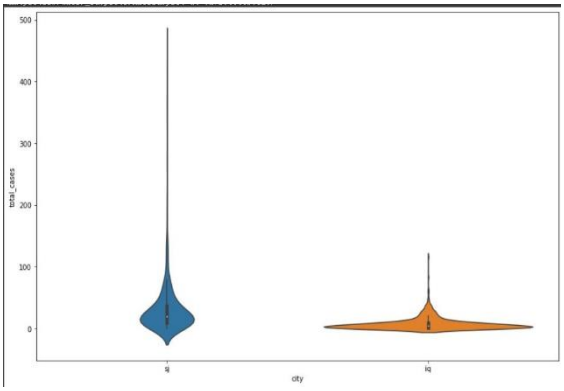


Figure 3a. Histogram of cases 3b. violin distribution of cases

- Difference of Features

From the plots of several features **Fig 4**, the values of same feature are quite different in two cities. Besides, there are no obvious change of the values of each feature correlated to time in the plots. In other word, they are randomly distributed.

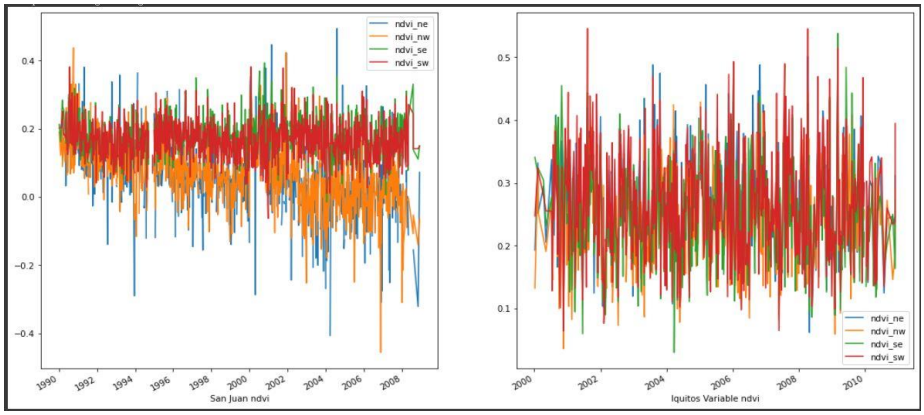


Figure 4a. The Plot of NDVI

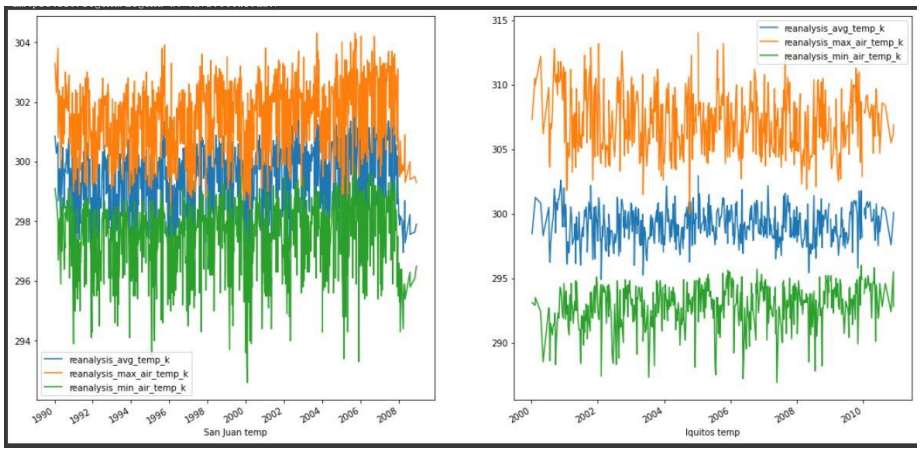


Figure 4b. The Plot of Temp

2.1.3 Correlation Analysis

From the lmplot of correlation between each feature and total cases **Fig 5**, the correlation is not significant in general. Besides, comparing to San Juan, the correlation in Iquitos is more obvious. In this case, the relation of features and target variable are closer to dependent in Iquitos, which may cause better predication in further steps.

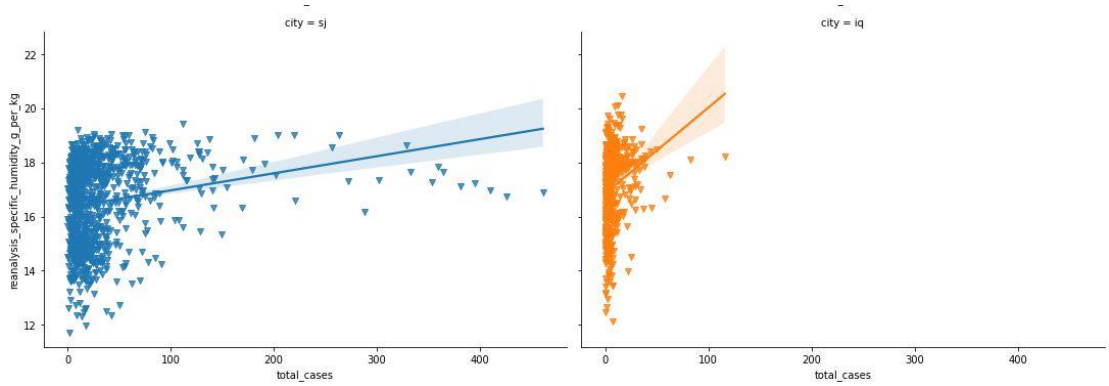


Figure 5a. LMplot of humidity and total cases

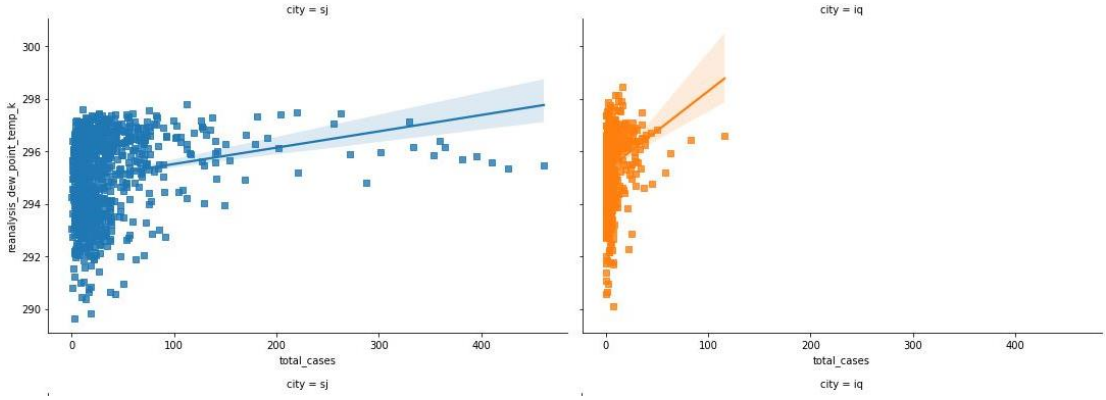


Figure 5b. LMplot of Tempure and total cases

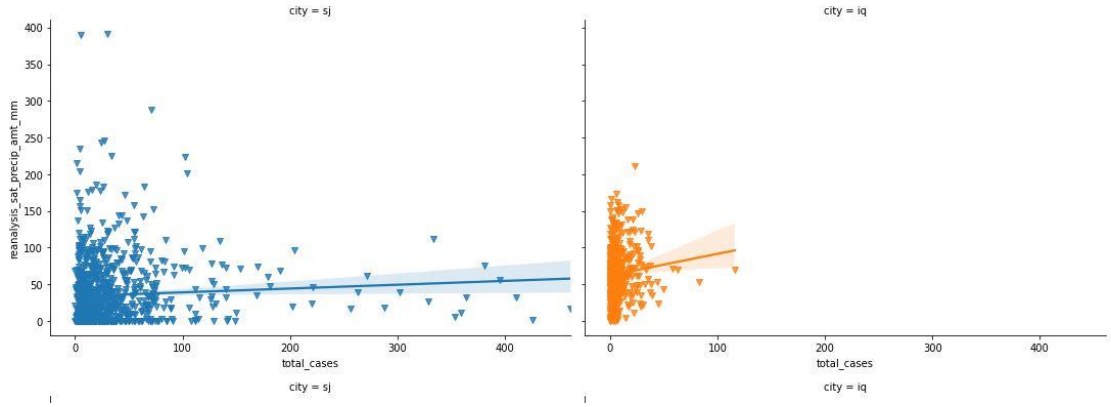


Figure 5c. LMplot of precipitation and total cases

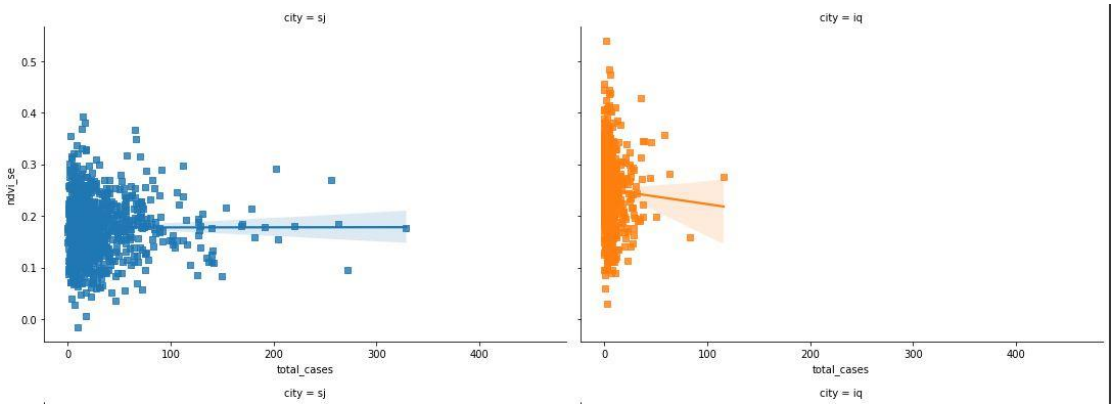


Figure 5d. LMplot of NDVI and total cases

2.2 Missing Data Imputation

There are lots of missing data inside this dataset **Fig 6**. After analyzing those missing data, different approaches should be applied to difference cases.

year	0
weekofyear	0
ndvi_ne	191
ndvi_nw	49
ndvi_se	19
ndvi_sw	19
reanalysis_air_temp_k	6
reanalysis_avg_temp_k	6
reanalysis_dew_point_temp_k	6
reanalysis_max_air_temp_k	6
station_avg_temp_c	6
station_max_temp_c	6
station_min_temp_c	6
reanalysis_min_air_temp_k	6
reanalysis_specific_humidity_g_per_kg	6
station_diur_temp_rng_c	6
precipitation_amt_mm	9
reanalysis_sat_precip_amt_mm	9
reanalysis_precip_amt_kg_per_m2	6
station_precip_mm	6
reanalysis_relative_humidity_percent	6
reanalysis_tdtr_k	6
total_cases	0
dtype: int64	

year	0
weekofyear	0
ndvi_ne	3
ndvi_nw	3
ndvi_se	3
ndvi_sw	3
reanalysis_air_temp_k	4
reanalysis_avg_temp_k	4
reanalysis_dew_point_temp_k	4
reanalysis_max_air_temp_k	4
station_avg_temp_c	37
station_max_temp_c	14
station_min_temp_c	8
reanalysis_min_air_temp_k	4
reanalysis_specific_humidity_g_per_kg	4
station_diur_temp_rng_c	37
precipitation_amt_mm	4
reanalysis_sat_precip_amt_mm	4
reanalysis_precip_amt_kg_per_m2	4
station_precip_mm	16
reanalysis_relative_humidity_percent	4
reanalysis_tdtr_k	4
total_cases	0
dtype: int64	

Figure 6a. Sum of Missing Data in San Juan 6b. Sum of Missing Data in Iquitos

- Extreme case

For the feature contains too mand massing values (account for 20%), as the mark in Fig 6a, the imputation approach should be carefully selected. The histogram shows the normal distribution of ndvi_ne in San Juan **Fig 7**. In this case, we could use global mean values to proceed imputation.

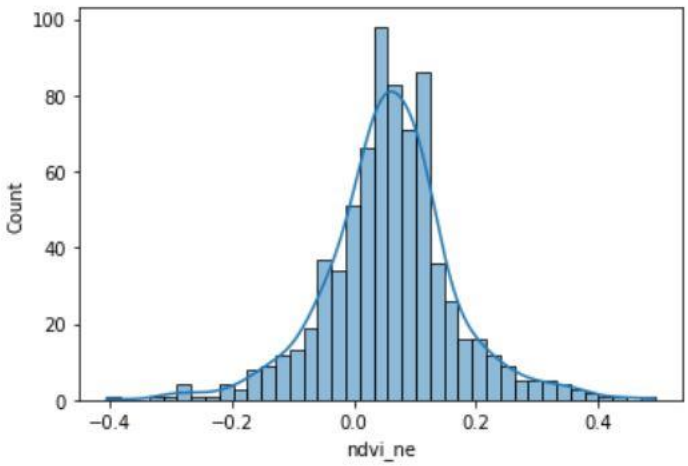


Figure 7. Histogram of ndvi_ne in San Juan

- Other cases

As for other features containing fewer missing values (less than 10%), the global median can by applied to impute missing values. Global median has less impact of outliers and thus this approach can be commonly used in the case of few missing data.

2.3 Feature analysis

Feature selection is the process of identifying and selecting a subset of input variables that are most relevant to the target variable. The goal of feature analysis id to identify some or all of those features that are relevant to the target. In other word, they are dependent to target variable. The other goal is to identify and remove some of the redundant input features, which means the features themselves should be independent to each other.

2.3.1 Feature Selection

The purpose of feature selection here is to eliminate highly dependent variables. The Pearson correlation coefficient are applied here to measure of linear correlation between two features. It is the covariance of two variables, divided by the product of their standard deviations, thus it is essentially a normalized measurement of the covariance, such that the result always has a value between -1 and 1.

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

From the marks in heat map **Fig 8**, there are several highly corelated variables ,which should be eliminated.

Figure 8a. Heatmap of Feature Correlations before Selection

Figure 8b. Heatmap of Feature Correlations after Selection

2.3.2 Feature Aggression

Feature Aggression is the process of combining two or more attributes into a single attribute. The purposes of Feature Aggression is to reduce the number of attributes or objects, to change of scale of dataset and to obtain more stable data because of less variability of aggregated data.

Given the heatmap after feature selection, there are still some high corelated area (marked in **Fig 8b**). To deal with this, feature aggression is applied here. Specifically, I aggregate four NDVI values with their means because of its normal distribution. I also aggregate the temperature features to reduce their reduction and correlation **Fig 9**.

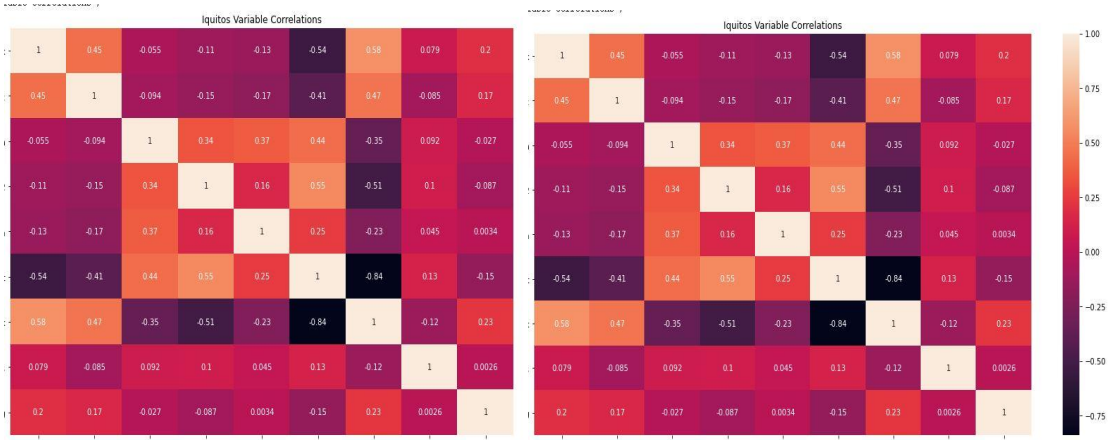


Figure 9. Heatmap after Feature Aggression

2.3.3 PAC

To compare with manually feature selection, I also applied PAC to original dataset. The Key idea of Principle Component Analysis (PCA) is to extract low dimensional set of features from a high dimensional data set (>=3) with a motivation to capture as much information as possible.

The theory of PCA is not complex. PCA defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by some scalar projection of the data comes to lie on the first coordinate as first principal component, the second greatest variance on the second coordinate as second principal component, and so on.

In this project, I firstly applied data normalization to original data, for which I will explain in more detail below. After that, I selected the number of principal components that account for 95% of information of original data Fig 10.

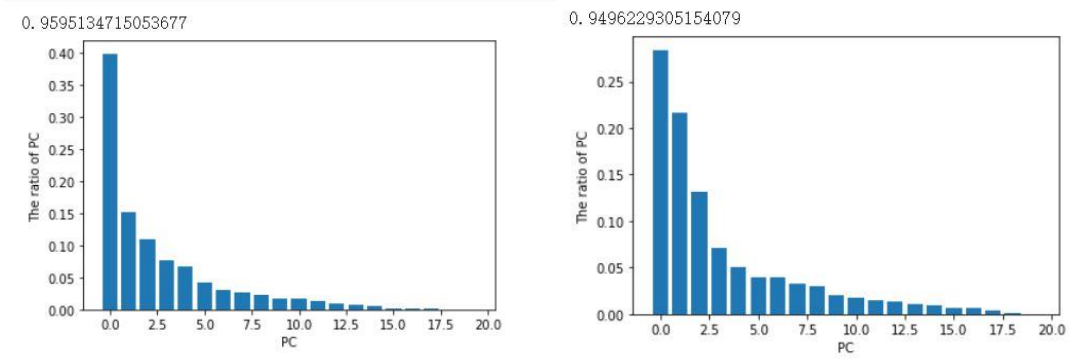


Figure 10a. PCA ratio of San Juan 10b. PCA ratio of Iquitos

2.4 Data Normalization and Split

2.4.1 Data Normalization

I compared two normalization method **Table 1**. From the boxplot of several features Fig 11, although the population of them are not totally normal distributed, however, some of features have lost of outliers which has impact more on the normalization. After considering this property of dataset, I chose z-score normalization to scale the data[2].

Table 1. Comparation of method of normalization

	Standard score	Min-max feature scaling
Formula	$\frac{x-\mu}{\sigma}$	$\frac{X-X_{min}}{X_{max}-X_{min}}$
Description	have a mean (μ) of 0 and standard deviation (σ) of 1 (unit variance)	rescales the values into a range of [0,1]
Application	population parameters are known, especially for population are normally distributed	all parameters need to have the same positive scale, yet sensitive to outliers

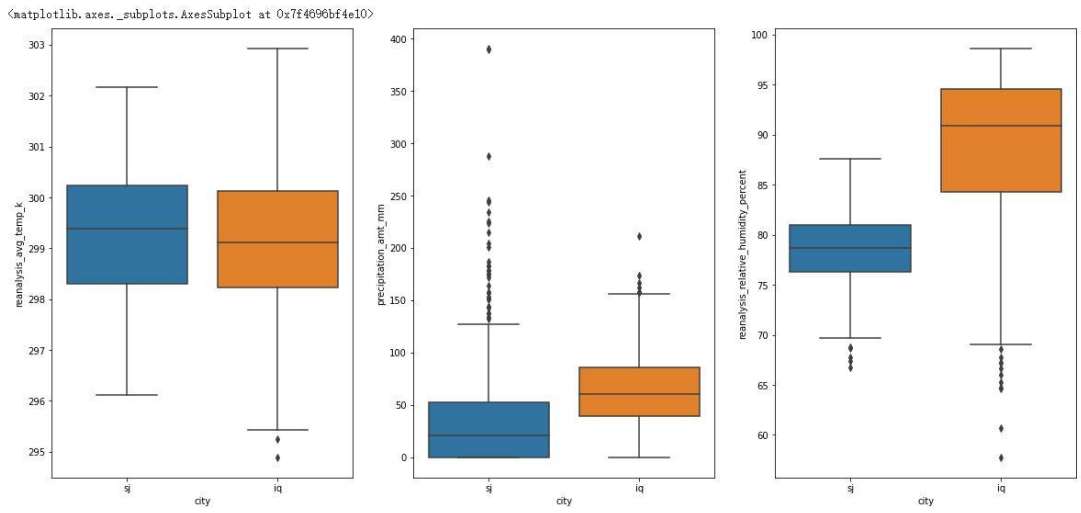


Fig 11. Box charts of features

2.4.2 Data Splitting

Since the data type is time series data, the property of time inside the data should be maintained to help prediction. Hence, normal data splitting method should not be applied here.

Here introduce a data spilt method called “TimeSeriesSplit”, which is in sklearn.model_selection. Time Series cross-validator provides train and test indices to split time series data samples that are observed at fixed time intervals. In each split, test indices must be higher than before, and thus shuffling in cross validator is inappropriate[2].
This cross-validation object is a variation of KFold. In the kth split, it returns first k folds as train set and the (k+1)th fold as test set. Unlike standard cross-validation methods, successive training sets are supersets of those that come before them.

For this project, the first 75% data are split for training, and the other 25% are regard as testing data.

2.5 Model Training

2.5.1 Liner regression

The first regression used here is Liner regression. In statistics, linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables. it attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.
Since the previous analysis already shows the target variable and the other variables are not linearly correlation, the liner regression can server as the baseline of regression model.

2.5.2 Random Forest Regression

Random forests are an ensemble learning method for machine learning tasks such as classification, regression. Random forests are operated by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or prediction of the individual trees[3]. Random decision forests correct for decision trees' habit of overfitting to their training set.

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The two most important parameters of these estimators are *n_estimatorsint* and *max_depthint*.

- *n_estimatorsint*: The number of trees in the forest.
- *max_depthint*: The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.

From the result of tunning process **Fig 12,** the parameters should be set as n_estimatorsint = 260, max_depthint = 10.

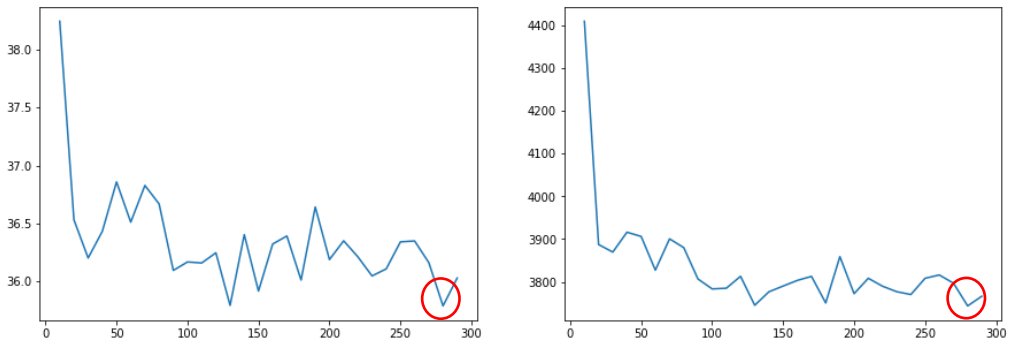
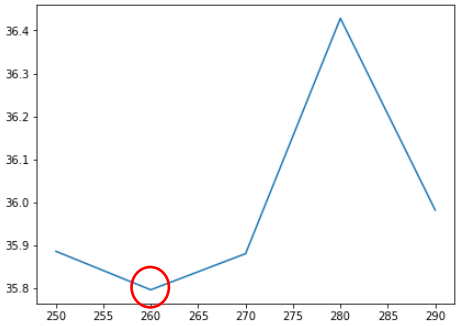


Fig 12a. MAE of different n_estimatorsint 12b. MSE of different n_estimatorsint



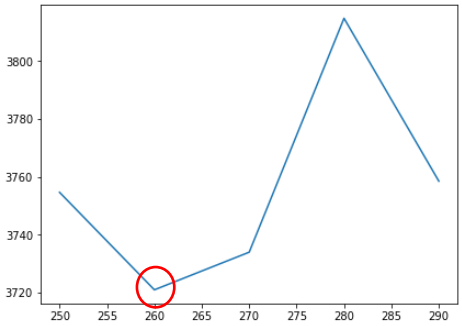
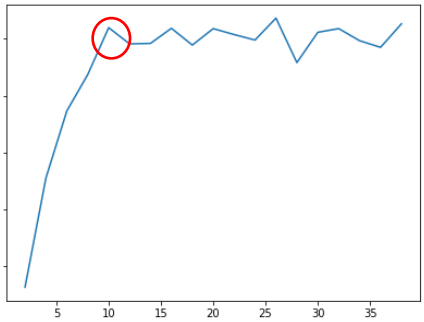


Fig 12c. MAE of different n_estimatorsint 12d. MSE of different n_estimatorsint



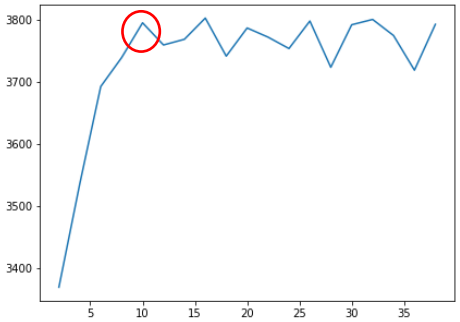


Fig 12c. MAE of different max_depthint 12d. MSE of different max_depthint

2.6.3 SVM_SVR

Support Vector Regression (SVR) is characterized by the use of kernels, sparse solution, and VC control of the margin and the number of support vectors. The kay idea of SVR is to decide a decision boundary at a distance from the original hyperplane such that data points closest to the hyperplane or the support vectors are within that boundary line **Fig 13.** Thus, SVR provides the flexibility to define how much error is acceptable in our model and will find an appropriate line to fit the data.

The objective function of SVR is to minimize the coefficients not the squared error. The error term is instead handled in the constraints, where we set the absolute error less than or equal to a specified margin. The model can tune epsilon to gain the desired accuracy. To minimize the coefficients:

$$MIN\frac{1}{2}\ ||w||^2$$

To constraint the epsilon error:

$$|y_i - w_i x_i| \leq \varepsilon$$

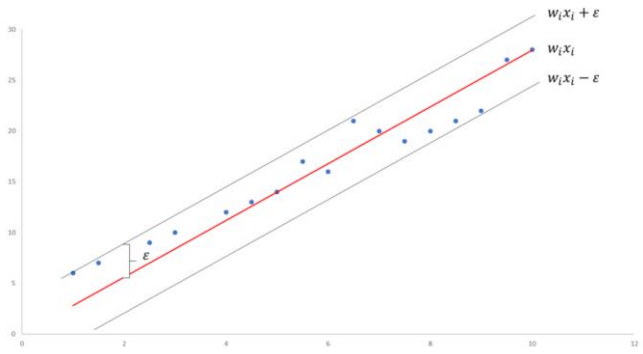


Fig 13. The schematic diagram of SVR

2.6.4 Ensemble model

- XGBoost

Boosting is a procedure to adaptively change distribution of training data by focusing more on previously misclassified records.

Like bagging, a weight for each training example is given initially, and all training examples are assigned equal weights. However, unlike bagging, in boosting, weights may change at the end of a boosting round. Besides, for ensemble, a weight is provided to each classifier based on its accuracy.

The Basic Idea of Boosting is that: 1) Examples that are wrongly classified will have their weights increased; 2) Examples that are classified correctly will have their weights decreased.

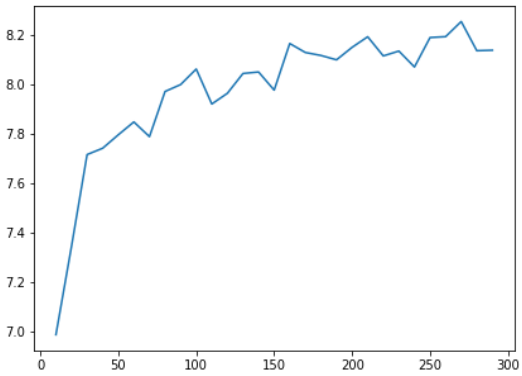
- Gradient Boosting Regressor

Gradient Tree Boosting or Gradient Boosted Decision Trees is a generalization of boosting to arbitrary differentiable loss functions. It is an accurate and effective off-the-shelf procedure that can be used for both regression and classification problems in a variety of areas including Web search ranking and ecology.

a. *n_estimators*: control the number of weak learners
b. *max_depth* : control the size of each tree to early stop the training
c. *max_leaf_nodes*: control the size of trees to early stop the training
d. *learning_rate*: control the overfitting via shrinkage

The two most important parameters of these estimators are *n_estimators* and *learning_rate* because of a trade-off between them. From the result of tunning process **Fig 14**, the parameters should be set as

n_estimatorsint = 10, learning_rate = 0.3.



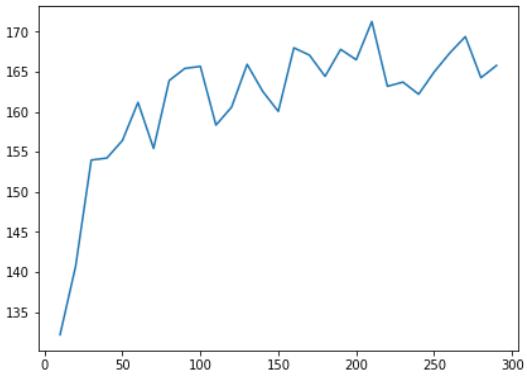
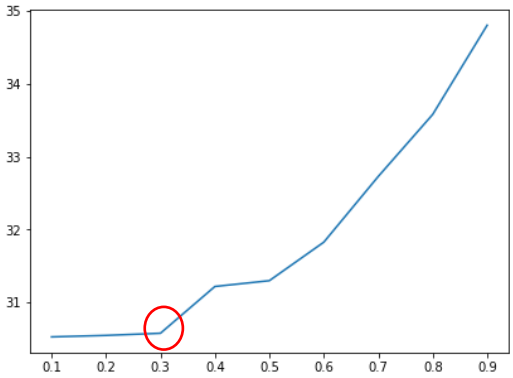


Fig 14a. MAE of different n_estimatorsint 14b. MSE of different n_estimatorsint



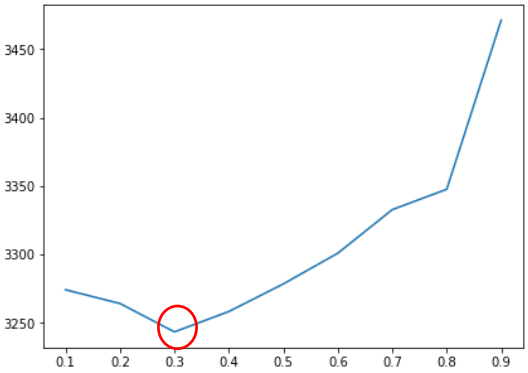


Fig 14c. MAE of different learning_rate 14d. MSE of different learning_rate

- Voting Regressor

The idea behind the Voting Regressor is to combine conceptually different machine learning regressors and return the average predicted values[4]. Such a regressor can be useful for a set of equally well performing models in order to balance out their individual weaknesses.

After comparing the performance of previous model, the SVR and Random Forest Regression are chosen in Voting Regression here.

3.Results and Analysis

3.1 Evaluation of Performances

Table 2. The Performance of Single Models

Model Name	Feature Selection	San Juan		Iquitos	
		MAE	MSE	MAE	MSE
Linear Regression	Aggression	27.365	1139.062	8.352	185.81
	PCA	19.31042	793.8808	9.23	171.7
Random Forest Regression	Aggression	29.443	1476.09	10.75	282.54
	PCA	16.40166	676.807	8.9357	170.465
SVR	Aggression	17.44	741.277	8.35	214.67
	PCA	16.22779	756.615	8.0472	180.5797
XGBoost	PCA	24.04273	1347.54	9.4923	243.0153
Gradient Boosting	PCA	18.3461	1034.714	8.36153	197.7769
Voting Regressor	PCA	18.67687	751.565	8.896161	171.4358

Table 2 shows all MAE score of all models in two cities. In both cities, the SVR and voting regression performs better with the higher MAE scores. Whereas the random forest regression and linear regression preforms worse, with lower scores Fig 15.

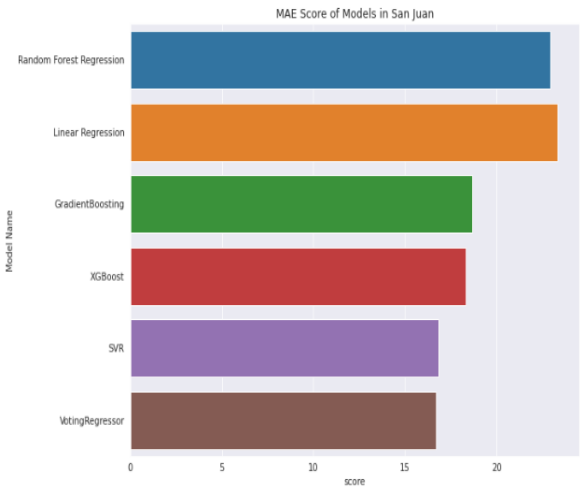
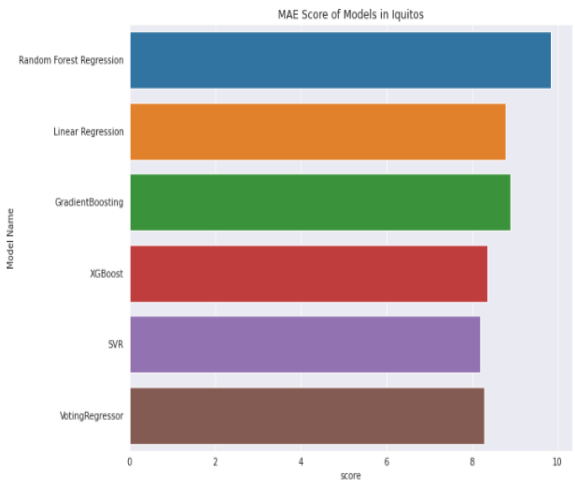


Fig 15a. MAE of Models for San Juan



15b. MAE of Models for Iquitos

3.2 Comparation and Analysis

3.2.1 Comparation of results in two cities

From the bar chart **Fig16**, the difference of scores between cities is obvious. Each Model of Iquitos preforms much better than San Juan. From the previous data exploration **Fig 5**, the reason may rely on the higher correlation between features and cases in Iquitos. Besides, from the plot of total cases **Fig 1a**, cases changing of San Juan are harsher than that in Iquitos. In this case, there may be reasons other than climate causes the large break of dengue in curtain period.

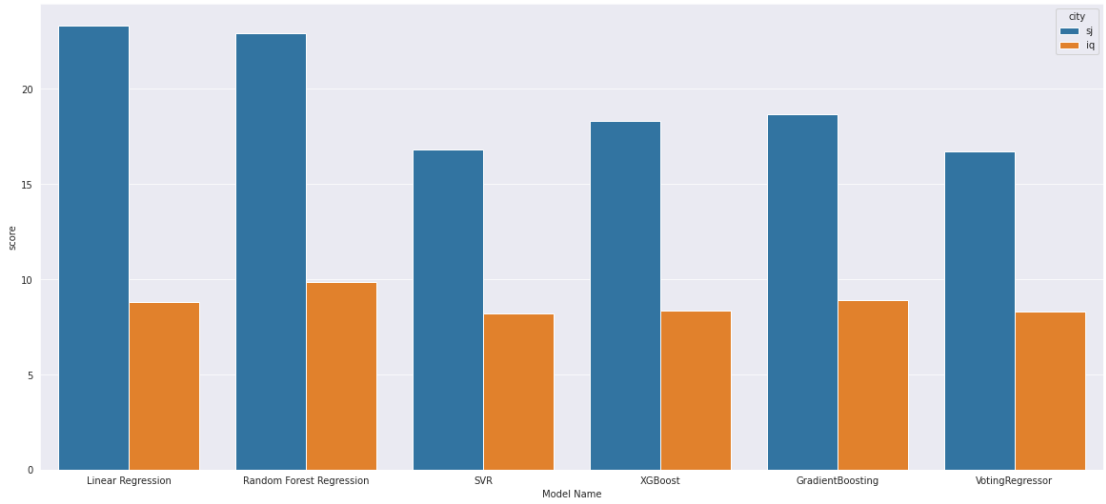


Fig 16. Comparation of MAE Scores between cities

3.2.2 Comparation of Data selection

From the bar plot, the method of PAC preforms better in each single model. Within that, the Random forest regression is more sensitive to the different data selection method. Whereas the SVR is the most robust.

The reason is that feature aggression in this project does not contain the weight of each features. To decide those weights, the more information besides dataset is in need. However, Random forest regression are consisted of the decision trees, and the

generation of each decision tree could not involved in all features. In other words, feature selection already processed by the algorithm of random forest. In this case, if some of the importance features are ignored or counteracted by other features, the Random forest regression may further process worse feature selection, thus cause the drop the performance.

On the other hand, the robust performance of SVR may also because of its algorithm. Both SVM_SVR and PCA involved in distance-based algorithm. When encounter high dimensions situation, the SVM could transform data into higher dimensional space, which is kinds of the opposite way of PCA. Even though the methods seems opposite at the first glance, the main ideas of those algorithm are similar, which may cause the close results of two feature selection methods in SVM_SVR.

However, the lower score shows in the feature aggression of linear regression in Iquitos, which is different from other models Fig 17. It is may cause of the previous strict correlation analysis and selection as well as the higher correlation itself inside the data of Iquitos. Both factors consist with the assumption of linear regression such that the feature aggression have lower score.

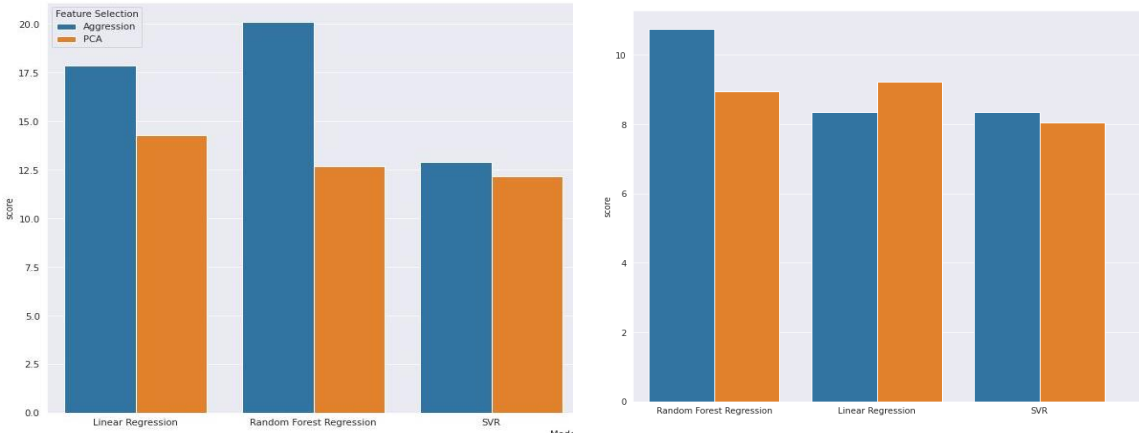


Fig 17. Comparation of MAE between Methods of Features Selection

3.3 Prediction visualization

The SVM_SVR model is chosen as best model to visualize the result of prediction **Fig 18.** From the line plot, the prediction of Iquitos is better than San Juan. However the overall predictions are still not promising. There are several factors may contribute to the results:

- The properties of dataset

According to the key idea of regression, features should be dependent to target variable, and yet the features themselves should be independent to each other. Regression analysis typically estimates the conditional expectation of the dependent variable given the independent variables.

However, after analysis, this dataset fails to meet both requirements. From the previous data exploration, the correlation between features and target variables are weak, which means features are not dependent to target variable. Besides, the similar climate data (e.g. all kinds of temperature values) collected from different sources are actually highly correlated to each other, which also indicates features are not independent to each other. All in all, From the relationship within features and the relationship between features and target variables, this dataset may not suitable for regression.

- The hardness of data process

The redundancy of climate features causes the hardness of data process, especially feature selection. There are only four types of data (temp, humidity, precipitation, ndvi), yet it consists of 20 features in original dataset. In this case, the feature selection is tricky and may need other indicators.

- Other potential factors

The extreme high cases in San Juan in certain year may indicate another factor. One of the reasonable explanations is that comparing to Iquitos, there are some other potential

factors in San Juan may cause dengue spread. For example, the number of natural enemies of mosquitoes, the people's awareness of prevention or the supply of equipment for mosquito's prevention.

Those potential factors are more direct factors than climates. To be more specific, if natural enemies of mosquitoes are larger in one area, the likelihood of dengue spread is low, no matter how climate contribute to culture mosquitoes.

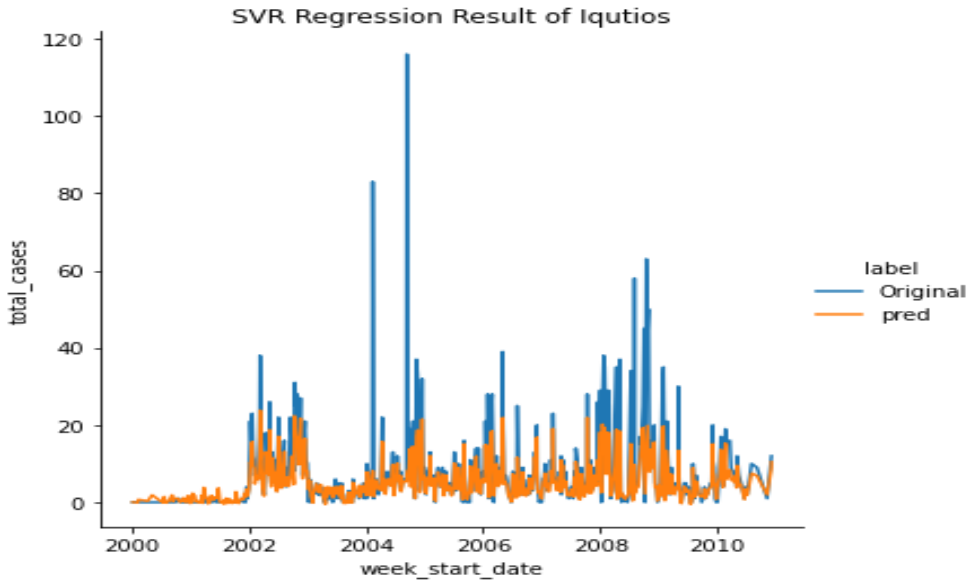


Fig 18a. SVR regression result of Iquitos

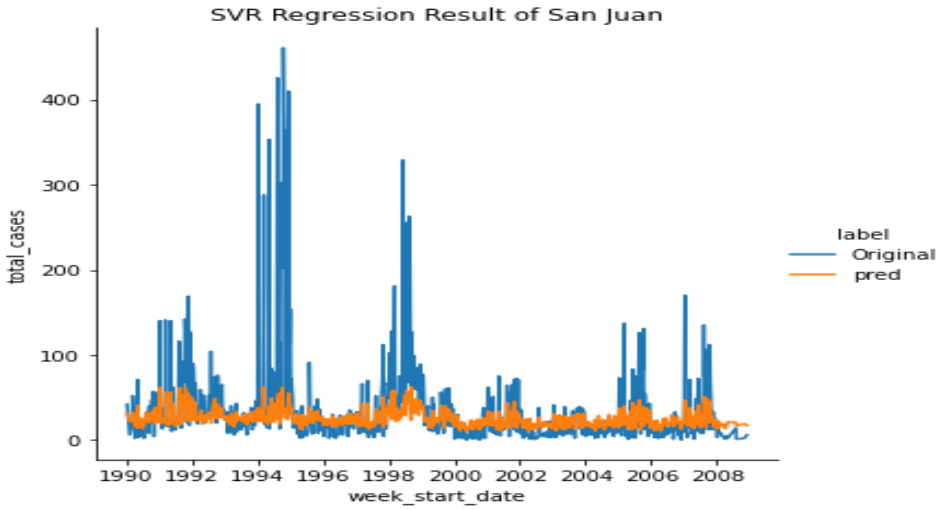


Fig 18a. SVR regression result of San Juan

4. Future Works

The basic knowledge of climate should be studied previously. With those knowledge, the more reasonable way of data process maybe raised and thus improves the performance of models.

The models used in this project are limited, there are many other models should be tried and evaluated, such as Poission regression, Bagging regression and RidgeCV. Those are also great models that may help to improve the performance of the models.

The classification should also be tried in this dataset. To be more specific, the label of total_cases can be divided into different classes and thus the models of classification can be applied to this dataset. The comparation of regression and classification helps to have a deep insight into the algorithms of machine learning.

5. Conclusion

Dengue fever is a mosquito-borne tropical disease caused of the dengue virus. The goal of this project is to predict the total_cases label given the time series dataset of climate in two cities.

After data analysis, data clean and features selection, multiple regression models are applied here to do prediction. Within them, the SVM_SVR model preforms best, with MAE score 16.2 in San Juan and 8.05 in Iquitos. Whereas, the Linear regression and Random Forest preform worse, with MAE score around 20 in San Juan and 9 in Iquitos.

The performance of models is compared through different variables. And the reasons of each difference are further discussed. The unpromising results may cause of various factors. The properties od dataset, the hardness of data process and other potential factors should also be taken into consideration.

References

[1] Méndez-Lázaro P, Muller-Karger FE, Otis D, McCarthy MJ, Peña-Orellana M. Assessing climate variability effects on dengue incidence in San Juan, Puerto Rico. Int J Environ Res Public Health. 2014 Sep 11;11(9):9409-28. doi: 10.3390/ijerph110909409. PMID: 25216253; PMCID: PMC4199026.

[2] Rob J Hyndman, with contributions from George Athanasopoulos, Slava Razbash, Drew Schmidt, Zhenyu Zhou, Yousaf Khan, Christoph Bergmeir, and Earo Wang. forecast: Forecasting functions for time series and linear models, 2014. URL http://CRAN.R-project.org/package=forecast. R package version 5.6.

[3] M Amasyali and O Ersoy. Classifier ensembles with the extended space forest. 2013.

[4] Terry Therneau, Beth Atkinson, and Brian Ripley. rpart: Recursive Partitioning and Regression Trees, 2014. URL http://CRAN.R-project.org/package=rpart. R package version 4.1-8.